

Universidad Autónoma de Occidente



Especialización Analítica de Big Data

Prof. Jennyfer Portilla Yela

Métodos Estadísticos Para el Análisis de Datos

Tarea Nro. 5

Sergio Alejandro López Caro

Emmanuel Alejandro Quiceno Rodríguez

Contenido

Introducción	3
Metodología.....	4
Resultados	5
Análisis exploratorio	5
Análisis de correlación	6
Modelo regresión lineal.....	8
Modelo de regresión lineal ajustado	8
Incursión de la variable Asia u otras en el modelo.	9
Conclusiones	10

Introducción

Dentro del presente informe se realiza la implementación de distintas herramientas estadísticas para el análisis del índice de felicidad en una muestra de distintos países. Dentro de este dataset es posible encontrar información de distintas variables como el nombre del país, su puntaje de GDP per cápita (también conocido como PIB), apoyo social, esperanza de vida saludable, libertad para tomar decisiones, generosidad y percepción de la corrupción.

El objetivo trazado para la realización de esta tarea será identificar las variables con un mayor impacto dentro de la felicidad (medida por un puntaje "Score") y lograr la construcción de un modelo de regresión lineal múltiple que permita caracterizar esta variable de felicidad. Dentro de este mismo dataset se realizará un análisis extra dentro del continente asiático, para realizar un análisis y comparación del comportamiento de esta variable.

A lo largo de este informe se podrá encontrar información respecto a análisis de correlaciones, ajuste y validación de modelos sumado a la comparación y conclusión de cada uno de los resultados obtenidos dentro de la exploración de los datos.

Metodología

Para lograr los objetivos propuestos dentro de este informe, se utiliza la herramienta R Studio con el fin de lograr la carga de la información del dataset, así como el análisis de este. Para iniciar, se realiza un análisis exploratorio de los datos para ver cómo se comportan, por ejemplo, es de interés conocer el valor de las medidas de tendencia central, las desviaciones estándar, y analizar el comportamiento por medio de gráficas (por ejemplo, histogramas). Por otro lado, también se analiza cuantos datos faltantes y nulos se presentan en el dataset.

Luego de realizar el análisis exploratorio de los datos, se procede a realizar un análisis de correlación por medio de matrices y gráficos que permiten visualizar e identificar si existe relación entre las diferentes variables del dataset, con el fin de ir evaluando si hay variables que pueden explicar a otras. Se realiza una evaluación por pares, es decir, se comparan de a dos variables y se calcula que tanto explica una variable a la otra, donde es importante resaltar que para que exista relación fuerte entre cada par de variables se debe buscar valores del coeficiente de correlación entre -1 y 1. Por otro lado, entre más cerca esté el coeficiente de correlación este del valor 0, menos relación existirá entre las variables.

Posteriormente, con el análisis de correlación realizado, se procede a estimar un modelo de regresión lineal múltiple, donde se incluyen todas las variables y se evalúa la calidad de este, con el objetivo de eliminar del modelo las variables que no sean relevantes para el modelo. Una vez eliminadas las variables menos significativas, se realiza un nuevo modelo de regresión con las variables existentes. Después, se ingresará al modelo una variable "Asia u otras", y se repetirá el paso anteriormente mencionado.

Por último, se realiza una comparación de los modelos realizados anteriormente y se concluye acerca de los resultados obtenidos.

Resultados

Análisis exploratorio

Al momento de realizar la carga de datos dentro de R estudio, es posible extraer información sumamente importante para el desarrollo del informe; gracias al análisis exploratorio de los datos es posible comprender la información que se obtiene del dataset, en donde se cuenta con 150 países y 10 variables. Estas variables están relacionadas con el nombre del país, su puntaje de felicidad, PIB, percepción de la corrupción, apoyo social, esperanza de vida saludable, libertad para tomar decisiones, generosidad y el continente al cual pertenecen y en caso de ser un país extenso en territorio, tendrá valores distintos en su continente, por lo que es posible identificar de entrada variables tanto numéricas como de texto.

Dentro de este dataset es posible observar que la variable score será la variable principal dentro de los datos, con un valor promedio de 5.40 con un rango entre 2.85 y 7.77 (Tabla 1), mientras que si se analiza las demás variables esas cuentan con una **dispersión** mayor, en donde se destacan mayoritariamente generosidad y percepción de la corrupción.

Variable	Mínimo	Q1	Mediana	Media	Q3	Máximo
Score	2.853	4.551	5.362	5.408	6.180	7.769
GDP per capita	0.000	0.586	0.954	0.898	1.228	1.684
Social support	0.000	1.055	1.268	1.205	1.442	1.624
Healthy life expectancy	0.000	0.539	0.789	0.722	0.881	1.141
Freedom to make choices	0.000	0.306	0.416	0.390	0.498	0.631
Generosity	0.000	0.110	0.177	0.185	0.247	0.566
Perceptions of corruption	0.000	0.047	0.083	0.106	0.139	0.453

Tabla 1. Descripción estadística de variables numéricas

Como parte de este análisis exploratorio de los datos, también se buscará analizar si estos tienen valores nulos o faltantes dentro de sus filas con el fin de saber si dado el contexto, se requerirá una eliminación o imputación en los datos; al hacer el análisis de esto se descubrió que el dataset dentro de sus datos numéricos no contiene ningún valor nulo. Posteriormente se realiza una variable cuyo contenido estará enfocado íntegramente en los valores numéricos para facilitar la manipulación de los datos y su

análisis estadístico. En la Figura 1, se observa el mapa de valores faltantes que corrobora la información mencionada anteriormente.

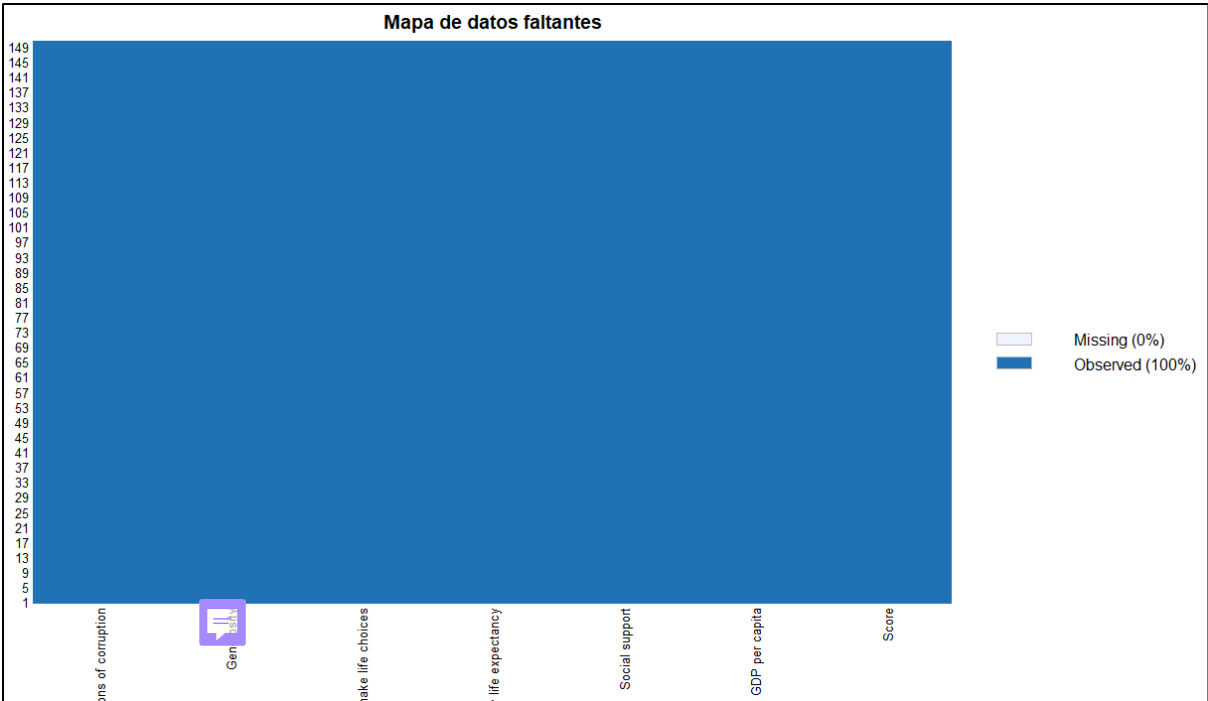


Figura 1. Mapa de datos faltantes.

Análisis de correlación

Para realizar el análisis de correlación se realiza un apoyo completamente en los recursos visuales obtenidos de R Studio y sus librerías, en este caso “corrplot” (Figura 2) y “ggally” (Figura 3).



Figura 2. Matriz correlación numérica (corrplot)

En la Figura 2 es posible observar el comportamiento de las variables al ser comparadas por medio de pares, donde es importante tener en cuenta que los valores cercanos a 0 significan que no existe una relación lineal entre las variables, mientras que los números próximos a 1 o -1 corresponden a una relación lineal positiva o negativa respectivamente. Se puede observar, por ejemplo, que las variables de “PIB”, “apoyo social” y “esperanza de vida saludable” presentan una relación lineal positiva con respecto al índice de felicidad, mientras que las variables “generosidad” y “percepción de la corrupción” parecieran no tener relación con este puntaje. Por otro lado, también se puede evidenciar que las variables “percepción de la corrupción”, “generosidad” y “libertad para tomar decisiones en la vida” no presentan una relación fuerte con ninguna de las otras variables, por lo cual se puede identificar posibles variables que serán eliminadas más adelante del modelo.

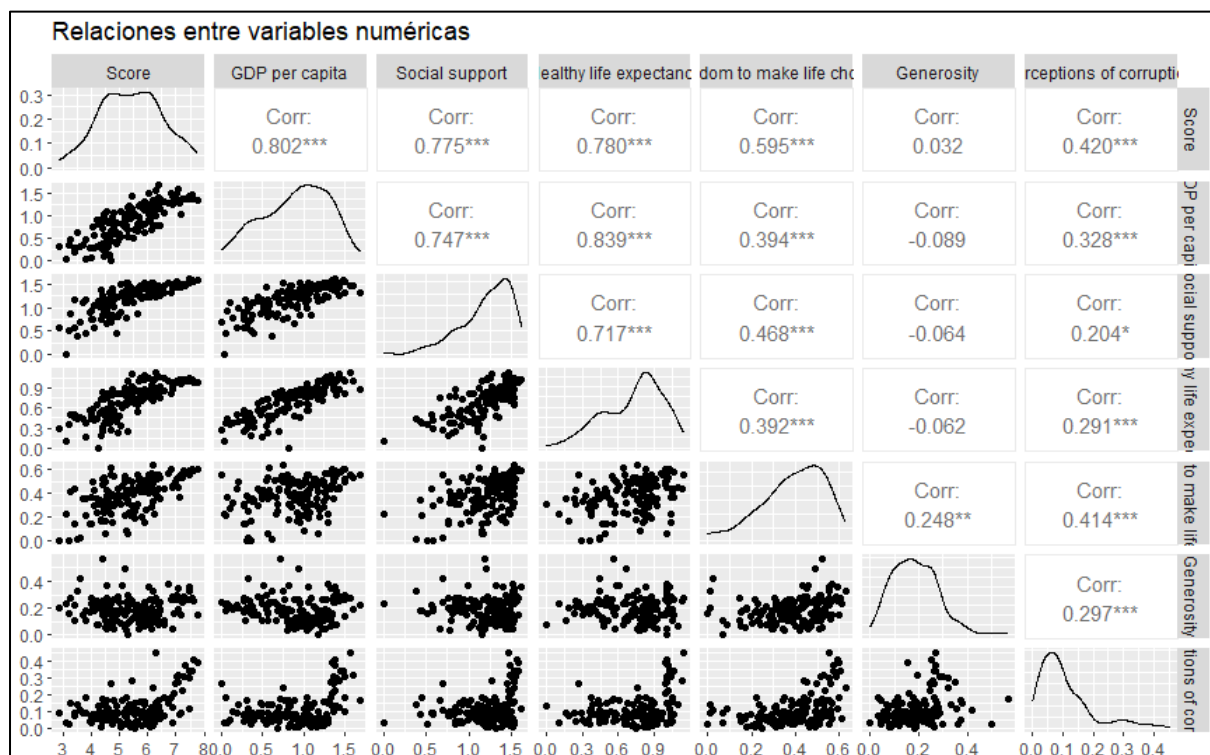


Figura 3. Matriz visual de correlación con dispersión

En la Figura 3, se puede observar mejor el análisis realizado anteriormente, ya que, gracias a este enfoque detallado ofrecido por la librería “ggally”, no solo es posible observar los valores numéricos de correlación entre variables, sino también se aprecia la forma y dispersión, en donde es posible visualizar el poco impacto que puede estar generando dentro del dataset las 2 últimas variables de “generosidad” y “percepción de corrupción”, las cuales deberán ser usadas más adelante para plantear un mejor modelo del dataset.



Modelo regresión lineal



Luego de realizar el análisis de correlación, se procede a estimar un modelo de regresión lineal múltiple, utilizando R Studio, obteniendo los resultados de la Tabla 2:

Variable	Estimación	p-valor
<i>(Intercepto)</i>	1,97	2E-16
<i>GDB per capita</i>	0,78	0,000322
<i>Social support</i>	1,012	0,0000156
<i>Healthy life expectancy</i>	0,997	0,002419
<i>Freedom to make life choices</i>	1,66	0,0000092
<i>Generosity</i>	0,0057	0,99
<i>Perceptions of corruption</i>	1,403	0,0113



Tabla 2. Modelo de regresión lineal múltiple.

De esta tabla es posible concluir que las variables “generosidad” y “percepción de la corrupción” no aportan valor al modelo, por lo cual, deberán ser eliminadas para plantear un nuevo modelo. Por otro lado, el valor del intercepto es 1.97, es decir, que si todas las variables del modelo tuvieran un valor de cero, el índice de felicidad sería de 1.97. También se puede observar que la variable “libertad para tomar decisiones” es la que más aporte al índice de felicidad tendrá. Sin embargo, esto puede significar valores residuales más altos si el modelo no está bien ajustado.

Modelo de regresión lineal ajustado

Luego de realizar el modelo de regresión lineal, se repite el mismo procedimiento eliminando en esta oportunidad las variables que no son relevantes para el modelo, es decir, aquellas que no están relacionadas con la variable de interés (índice de felicidad). En la Tabla 3, se observan los resultados obtenidos.

Variable	Estimación
<i>(Intercepto)</i>	2,0069
<i>GDB per capita</i>	0,8793
<i>Social support</i>	0,9055
<i>Healthy life expectancy</i>	1,0224
<i>Freedom to make life choices</i>	1,9991

Tabla 3. Modelo de regresión lineal ajustado.

Para evaluar el modelo de regresión, se realiza la verificación de la homocedasticidad (variabilidad de los residuos es constante) y se analiza la normalidad de los residuos, donde se puede observar en la Figura 4, que los residuos no tienen varianza

constante (heterocedasticidad), por otro lado, se puede observar que estos si presentan un comportamiento normal.

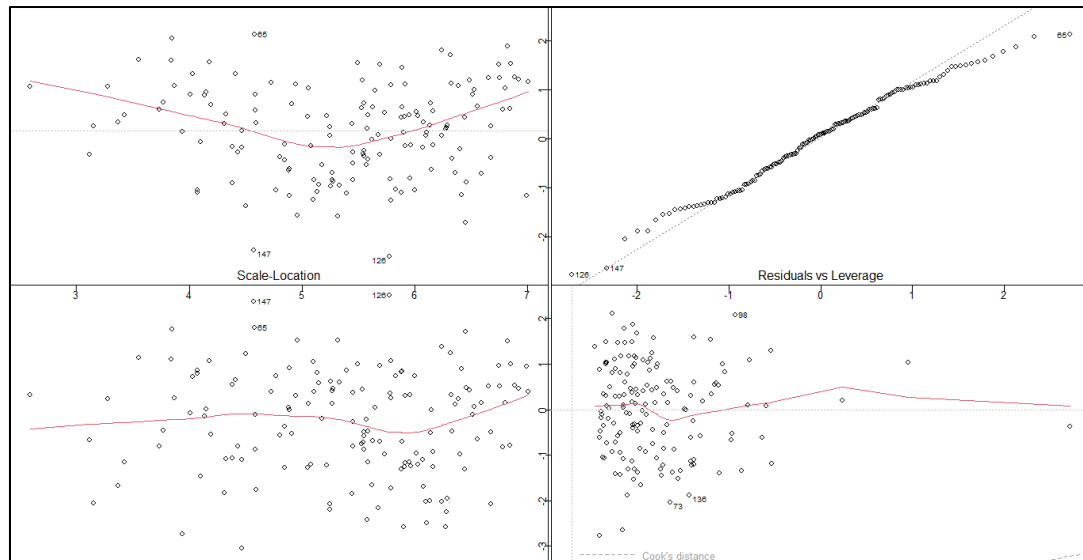



Figura 4. Supuestos del modelo de regresión.

Por último, se calcula el valor del coeficiente de determinación, obteniendo un valor de 0.7788, es decir, el modelo explica en un 77.88% la variabilidad de los datos. Es importante destacar que si  entre más cercano a 100% este el coeficiente mejor se explica la variabilidad de los datos, no siempre esto indica que el modelo sea bueno.

Incursión de la variable Asia u otras en el modelo.

Luego de realizar el modelo mencionado anteriormente, se realiza la incursión en el programa R Studio de la variable creada “Asia u otros” y se vuelve a estimar el modelo de regresión lineal, sin las variables que fueron eliminadas (generosidad y percepción de la corrupción), **obteniendo los mismos resultados analizados anteriormente.**



Conclusiones

- El análisis exploratorio de los datos permite caracterizar las variables objeto de estudio con el fin de evaluar sus medidas de tendencia central y su variabilidad. Este paso es muy importante para cualquier investigación ya que es el punto de partida para conocer las variables inicialmente, realizar estimaciones sin una visualización preliminar de los datos puede llevar a una mala interpretación de los resultados.
- Las variables “GDB per capita”, “Apoyo social” y “Esperanza de vida saludable” son los principales factores que influyen en el índice de felicidad de los países.
- Las variables “Generosidad” y “Percepción de la corrupción” no tienen un impacto significativo en el índice de felicidad.
- La inclusión de la variable “Asia” en el modelo no mejoró ni empeoró el modelo, por lo que se puede concluir que la felicidad no está directamente relacionada con la ubicación geográfica, sino con factores socioeconómicos de los países.
- El modelo explica en un 77.88% la variabilidad de los datos del índice de felicidad.
- Los residuos del modelo siguen aproximadamente una distribución normal.